Big Data Examination

Roll No. - DS5B-2121

Install PySpark

In [1]:
```
pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publ
ic/simple/
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 34 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 49.8 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642
sha256=550424d4405822124fd616b8ecfb7164d4c66f2df850b1516899280ad03053f9
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d53
34db05f2e83bcef74f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

Create Spark Context

Main entry point for Spark functionality. A SparkContext represents the connection to a Spark cluster, and can be used to create RDD and broadcast variables on that cluster.

In [ ]:
```
from pyspark import SparkConf, SparkContext
```

In [ ]:
```
if __name__ == "__main__":
  conf = SparkConf().setAppName("exam").setMaster("local[2]")
  # AppName : A name for your job, to display on the cluster web UI and Master: Cluster
  sc = SparkContext(conf = conf)
```

Create a RDD by reading the csv file

A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel.

In [ ]:
```
data = sc.textFile("amazon.csv")
```

Print the top five raws of dataset

In [ ]:
```
data.take(5)
```

Count total number of raw in dataset

In [ ]:
```
data.count()
```

Check the number of partitions

In [ ]:

```
data.getNumPartitions()
```

Out[ ]:

```
2
```

My Roll No. Is : DS5B-2137

Remove Heading

we remove heading so that we don't get any error while applying any rdd operation regarding heading raw

In [ ]:

```
header = data.first()
header
```

Out[ ]:

```
'product name,asin,product url,brand name,image url,mrp,sale price,discount percentage,p
roduct description,date first available,number of reviews,seller name'
```

In [ ]:

```
newdata = data.filter(lambda x: x!=header)
```

In [ ]:

```
newdata.take(4)
```

Out[ ]:

```
['BHAGIRATH Standard comfort Disposable Pollution/Surgical Elastic Mask Surgical Face Ma
sk with Earloop Great for Air Pollution Virus. 1,B085GL266P,https://www.amazon.in/BHAGIR
ATH-Standard-Disposable-Pollution-Surgical/dp/B085GL266P/,Bhagirath fab,https://images-n
a.ssl-images-amazon.com/images/I/61XinzsWunL._SL1100_.jpg, 16999.00,199.00,99%,IND mask
covers the user’s nose and mouth and provides a physical barrier to fluids and particula
te materials. The surgical masks referenced in this guidance document include masks that
are labeled as a surgical  laser  isolation  dental or medical procedure masks with or w
ithout a face shield.,2020-03-01 00:00:00,0,Bhagirath fab',
 'SHOPPERMART urgical Face Mask Disposable - Pack of 5,B07YPSZK39,https://www.amazon.in/
SHOPPERMART-urgical-Face-Mask-Disposable/dp/B07YPSZK39/,SHOPPERMART,https://images-na.ss
l-images-amazon.com/images/I/61BgE1t%2BUJL._SL1000_.jpg,,,,Easy breathable  effective mu
ltiplayer bacterial filter  soft  lightweight  comfortable and easy to wear  used in med
ical  dental  laboratory  food sectors  school  household  industry and multi-purpose  b
acteria filtration efficiency 99%  particle filtration efficiency (0.1 micron) 99%  viru
s filtration efficiency (0.1 micron) 99%  fluid repellent 120 mmhg  breathable < 49 0 pa
fluid resistant hygienic & skin friendly ,2019-10-03 00:00:00,2,',
 'ShopyBucket Standard 3 PCS comfort Disposable Pollution Elastic Mask Disposable 3 Ply
Face Mask with Earloop Great for Air Pollution Virus Protection & Personal Health Face M
ask,B0855V7MQ8,https://www.amazon.in/ShopyBucket-Standard-Disposable-Pollution-Protectio
n/dp/B0855V7MQ8/,Shopy,https://images-na.ssl-images-amazon.com/images/I/61SPmUgDP9L._SL1
100_.jpg, 798.00,510.00,36%,,2020-02-26 00:00:00,0,BasicDeal',
 "Ivaan Disposable Earloop Medical Face Masks Two Layer Non-Woven Pack Of 100 pcs,B082HD
NJCP,https://www.amazon.in/Ivaan-Disposable-Earloop-Medical-Non-Woven/dp/B082HDNJCP/,IVA
AN,https://images-na.ssl-images-amazon.com/images/I/71MXY9ud8OL._SL1500_.jpg,,,,Facemask
s help limit the spread of germs. When someone talks  coughs  or sneezes they may releas
e tiny drops into the air that can infect others. If someone is ill face masks can reduc
e the number of germs that the wearer releases and can protect other people from becomin
g sick. A face mask also protects the wearer's nose and mouth from splashes or sprays of
body fluids. It is ideal for every house in day to day activities as well as doctors  su
rgeons  dentist  dental assistance  nurses  landscapers  contractors  plumbers  extermin
ators  nail technicians and many more.,2019-12-08 00:00:00,6,"]
```

```
# This is formatted as code
```

Function 1

In [ ]:

```
def fun2(x):
  a = x.split(",")[-6]
  if a == '':
    return 0
  else:
    return float(a)
```

In [ ]:

```
newdata.map(fun2).max()
```

Out[ ]:

```
22050.0
```

Function 2

Filter product name , product brand, product url, product image url

In [ ]:

```
def fun1(x):
  a = x.split(",")
  return a[0], a[3], a[2], a[4]
```

In [ ]:

```
newdata.map(fun1).take(4)
```

Out[ ]:

```
[('BHAGIRATH Standard comfort Disposable Pollution/Surgical Elastic Mask Surgical Face M
ask with Earloop Great for Air Pollution Virus. 1',
  'Bhagirath fab',
  'https://www.amazon.in/BHAGIRATH-Standard-Disposable-Pollution-Surgical/dp/B085GL266
P/',
  'https://images-na.ssl-images-amazon.com/images/I/61XinzsWunL._SL1100_.jpg'),
 ('SHOPPERMART urgical Face Mask Disposable - Pack of 5',
  'SHOPPERMART',
  'https://www.amazon.in/SHOPPERMART-urgical-Face-Mask-Disposable/dp/B07YPSZK39/',
  'https://images-na.ssl-images-amazon.com/images/I/61BgE1t%2BUJL._SL1000_.jpg'),
 ('ShopyBucket Standard 3 PCS comfort Disposable Pollution Elastic Mask Disposable 3 Ply
Face Mask with Earloop Great for Air Pollution Virus Protection & Personal Health Face M
ask',
  'Shopy',
  'https://www.amazon.in/ShopyBucket-Standard-Disposable-Pollution-Protection/dp/B0855V7
MQ8/',
  'https://images-na.ssl-images-amazon.com/images/I/61SPmUgDP9L._SL1100_.jpg'),
 ('Ivaan Disposable Earloop Medical Face Masks Two Layer Non-Woven Pack 0f 100 pcs',
  'IVAAN',
  'https://www.amazon.in/Ivaan-Disposable-Earloop-Medical-Non-Woven/dp/B082HDNJCP/',
  'https://images-na.ssl-images-amazon.com/images/I/71MXY9ud8OL._SL1500_.jpg')]
```

To find the max sales price

Function 3

To find the minimum MRP

In [ ]:

```
def fun2(x):
  a = x.split(",")[6]
  if a == '':
    return "null"
  else:
    return float(a)
```

In [ ]:

```
newdata.map(fun2).filter(lambda x: x != "null").min()
```

```
Out[ ]: 45.0
```

Function 4

```
In [ ]: rdd1, rdd2 = newdata.randomSplit([2, 3], 17)
```

```
In [ ]: rdd1.count()
```

```
Out[ ]: 212
```

```
In [ ]: rdd2.count()
```

```
Out[ ]: 319
```

Function5

```
In [1]: unionresult = rdd2.union(rdd1)
        unionresult.collect()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-d8098455b1b8> in <module>()
----> 1 unionresult = rdd2.union(rdd1)

NameError: name 'rdd2' is not defined
```

Function 6

```
In [ ]: intersectresult = rdd1.intersection(rdd2)
        intersectresult.collect()
```

Function 7

```
In [ ]: new = carresult.collect()
        new
```

Function 8

```
In [ ]: subresult = rdd1.subtract(rdd2)
        subresult.collect()
```